View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by Queensland University of Technology ePrints Archive

Application of cross validation techniques for modelling construction costs during the very early design stage

Franco K T Cheung[1] and Martin Skitmore[2]
[1]Department of Building and Construction, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
[2]School of Urban Development, Queensland University of Technology, Brisbane Australia

Corresponding Author:
Franco K T Cheung
Department of Building and Construction, City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong,
bcfranco@cityu.edu.hk

19 August 2005 (version 3a)

This is author version of paper published as:

Cheung, Franco K.T. and Skitmore, Martin
(2006) Application of cross validation techniques for modelling construction costs during the very early design stage. Building and Environment 41(12):pp. 1973-1990.

Copyright 2006 Elsevier

Application of cross validation techniques for modelling construction costs during the very early design stage

Abstract

Building client/owners need estimates of likely construction costs for budgeting purposes early in the procurement process when little detailed design information is available beyond the type, size and location of the facility. One of the more sophisticated techniques available for this purpose is the Storey Enclosure Method, developed by James in 1954. This uses the basic physical measurements of the building envelope, together with an arbitrary set of multipliers, or weights, to forecast tender/bid prices. Although seldom used in practice, James succeeded in showing his method to be capable of significantly outperforming alternative approaches.

The research reported in this paper aimed firstly to reassess James' claims with new data and secondly to advance his method by using regression techniques to obtain the weights involved. Based on data from 138 completed Hong Kong projects for four types of building, two types of regression models were developed. This involved the use of sophisticated features such as leave-one-out cross validation to simulate the way in which forecasts are produced in practice and a dual stepwise selection strategy that enhances the chance of identifying the best model. An algorithm was also designed to select the appropriate parametric and non-parametric tests for objective and rigorous model evaluation against alternatives.

The results indicate that, contrary to James' claim, both his original method and the two regression-based alternatives are not significantly better or worse than other models. Surprisingly, the widely used floor area model was found to under-perform in terms of consistency for offices and private housing. For private housing in particular, it was felt that the Storey Enclosure Method was likely to offer good prospects of improvement on those methods currently in use in practice.

Keywords: Price forecasting, early stage, Storey Enclosure Method, cost model, cross validation, forecasting accuracy.

1. Introduction

In the feasibility (or early sketch design) stage of a building project, the freedom to modify the scope, requirements, standards and design is very high. Although design information available at this stage is usually very coarse and limited, construction clients are generally eager to know the likely building price, i.e. the lowest tender price, for budgeting purposes. Conventionally, practicing forecasters use the Floor Area Method (FAM), by which the total useable floor area is measured from the available sketch drawings and multiplied by a suitable rate drawn either from the forecaster's experience or a database of rates from previous projects or a mixture of the two. As FAM involves only one variable – the floor area - it is often termed a *single rate model* and belongs to a family of such models.

Another member of the single rate family is the now obsolete Cube Method, which, as the name suggests, uses a measure of the volume of the building as its single variable. Yet another, and perhaps the most sophisticated of the single rate models is the Storey Enclosure Model (JSEM) [1] developed by James' in 1954. Although seldom, if ever used in practice, JSEM attempts to incorporate the effects of the physical shape of the building, total floor area, vertical positioning of the floor area, storey heights and usable floor area below ground level, e.g. basements.

The only variable in JSEM is the total weighted enclosure area, or *storey enclosure area*. This is obtained by multiplying each of the areas involved by an arbitrary weighting factor (prescribed by James) aimed at reflecting the likely additional associated construction costs. So, for example, basement floors are given a greater weighting than ground floors due to the extra costs in excavation, etc involved in providing basement floors. The floor area for each storey, total external wall area, basement wall area and roof area are weighted in this way and summed to provide a total storey enclosure area. As with all single rate forecasting methods, the storey enclosure area is then multiplied by a suitable rate to provide the forecasted building price.

As JSEM quantifies different components of a building in its single rate variable, it arguably makes use of more information than FAM and therefore is expected to be more accurate. Indeed, James was able to show that this was the case, with an analysis of a range of buildings completed at the time of his work. Hence, JSEM was chosen for further development in the research reported in this paper. Instead of determining the relationships between the dimensions of building components and the building price

intuitively, regression analysis is used for model exploration. Cross validation was used for the model building process.

Cost analyses from 138 completed Hong Kong projects of offices, private housing, nursing homes, and schools were used. In addition to the use of cross validation to simulate the way in which forecasts are produced in practice, a dual stepwise selection strategy was also used to enhance the chance of identifying the best model. Two types of regressed models were generated from different candidate sets, the Regressed Model for James' Storey Enclosure Method (RJSEM) and Regressed Model for Advanced Storey Enclosure Method (RASEM). These new models were evaluated against the three other single price models, i.e. the original JSEM, FAM and cube models, in terms of their accuracy. To do this, an algorithm for selecting the appropriate tests for the comparisons was also designed and which includes three steps: (1) to measure the forecasting accuracy in terms of bias and consistency; (2) to compare the forecasting accuracy of these models by the use of different parametric and non-parametric tests and (3) to group the models that show the same potency together.

2. Major directions of model development

A variety of applications of regression analysis in the forecasting of building costs and prices have been developed since the mid 1970s. It was first used to model building prices for offices [2-6], schools [7], houses [8-10], homes for old people [11], lifts [12], electrical services [12], motorway drainage [13] and a few other types of building [14]. It was then used to model the prices of reinforced concrete frames [15, 16] and building services [17]. It has also been used to model the prices of components such as the beams of suspended-roof steel structures [18]. Since the mid-80s, a greater variety of mathematical techniques, such as probabilistic simulation [19-24], neural network [25-29] and fuzzy logic [29, 30] have been used.

Of these, very few have been taken up in practice, with the use of conventional (traditional) techniques outweighing all the other techniques [31-33]. The reason, it is suggested, may be due to many practising forecasters not being well-equipped enough to understand and use other, more elaborate, models (Fortune and Lee's [31]).

A more likely alternative is that there is little conclusive evidence of the superiority of any of the nontraditional models, with the demand for a move to a more scientific basis for forecasting coming mainly from academia, rather than practice [34, 35]. For new

models to be used, practitioners will need to be convinced that the benefits will exceed the costs involved. This implies the need for a logical and systematic approach to performance measurement and model evaluation.

3. Problem definition

The FAM, cube model and JSEM can be represented mathematically by:

$$P = \left(\sum_{i=0}^{m+n} f_i \right) \cdot R \tag{1}$$

$$P = \left(\sum_{i=0}^{m+n} f_i \cdot s_i \right) \cdot R \tag{2}$$

$$P = \left(\sum_{i=0}^{n} (2 + 0.15i) f_i + \sum_{i=0}^{n} p_i s_i + 2 \sum_{j=0}^{m} f'_j + 2.5 \sum_{j=0}^{m} p'_j s'_j + r \right) \cdot R \,, \tag{3}$$

where P is the forecasted price, f_i is the floor area at i storeys above ground, p_i is the perimeter of the external wall at i storeys above ground, s_i is the storey height at i storeys above ground, n is the total number of storeys above ground level, m is the total number of storeys below ground, f'_j is the floor area at j storeys below ground level, p'_j is the perimeter of the external wall at j storeys below ground level, s'_j is the storey height at j storeys below ground level, r is the roof area and R is the unit rate (determined by historical data).

To apply the JSEM to a high-rise building comprising a podium and a tower with repetitive floors commonly found in Hong Kong as well as other big cities in the world, JSEM can be simplified to avoid laborious measurement of areas. Given that $\sum_{i=0}^{n} p_i s_i = n p_{pt} s_{pt}$, where p_{pt} is the average perimeter of the superstructure and s_{pt}, is the average storey height of the podium; $\sum_{j=0}^{m} p'_j s'_j = m p_b s_b$, where p_b is the average perimeter of the basement and s_b, is the average storey height of the basement; $\sum_{j=0}^{m} f'_j = m f_b$, where f_b is the average floor area per storey for floors at basement level and $f'_0 \approx f'_1 \approx \ldots \approx f'_m \approx f_b$ (the floor area for each level of the basement is more or less the same, and is approximately equal to f_b); and $n = a + b$, where a is the number of storeys of the

podium and b is the number of storeys of the tower such that $f_0 \approx f_1 \approx \ldots \approx f_a \approx f_p$ (the floor area for each level of the podium is more or less the same, and is approximately equal to f_p), where f_p is the average storey area for floors at the podium level and $f_{a+1} \approx f_{a+2}, \ldots, f_b \approx f_t$ (the floor area for each level of the tower is more or less the same, and is approximately equal to f_t), where f_t is the average storey area for floors at tower level, JSEM is simplified to:

$$P = \left(2 - \frac{0.15}{2}\right) a f_p R + \frac{0.15}{2} a^2 f_p R + \left(2 - \frac{0.15}{2}\right) b f_t R + \frac{0.15}{2} b^2 f_t R + 0.15 a b f_t R + rR + n p_{pt} s_{pt} R + 2 m f_b R + 2.5 m p_b s_b R \quad (4)$$

Consider a high-rise building that has no podium, or that the average storey area for the podium is approximately equal to that of the tower, i.e., $f_p \approx f_t \approx f_{pt}$, where f_{pt} is the average storey area for floors above ground level, and $a + b = n$. The simplified equation becomes:

$$P = \left(2 - \frac{0.15}{2}\right) n f_{pt} R + \frac{0.15}{2} n^2 f_{pt} R + rR + n p_{pt} s_{pt} R + 2 m f_b R + 2.5 m p_b s_b R \quad (5)$$

The structure of Equations (4) and (5) clearly suggest the model building process to be a typical multiple linear regression problem. To identify the predictors for best subset models, the variables used in JSEM together with additional variables, such as the number of storeys, the square of the number of storeys and their interaction with storey height, were considered as a set of candidate independent variables. The unit rate '*R*' was excluded, because the tender price is not measured on a unit area basis in regressed models. Table 1 shows a full list of the candidate variables for the regressed models for buildings with and without basements.

4. Data

The cost analyses of four types of projects, offices (42 projects), private housing (50 projects), nursing homes (23 projects), and schools (23 projects), from a ten-year period, the 3rd quarter of 1988 to 2nd quarter of 1997, were chosen to be the data source. They were provided by one very established surveying practice in Hong Kong. This single data source approach ensures that the generated models are applicable as

practicing forecasters mainly rely on in-house data for forecasting. The data that were used for modelling are tabulated in Appendix B (enclosing Tables B-1 to B-4).

5. Model Building

A non-parametric approach that is based on the mean square error (MSQ) is adopted. It is chosen rather than a parametric approach because the statistical assumptions under a parametric approach may not be easily satisfied [36] due to the small sample sizes in this study which may cause the parametric estimate of the error rates to be biased [37]. In modelling, the termination criterion is to minimise the MSQ.

6. Reliability analysis

Figure 1 shows the framework used for the identification, selection and validation of the price models. A resampling method, leave-one-out cross validation, is adopted to select variables and evaluate models. This was chosen both for its intuitive appeal, as each error value can be thought of as a real error that may arise in the practice of forecasting [38] and also because cross validation is known to be markedly superior with small data sets [39].

The accuracy of statistical inference in the leave-one-out method is preserved by dividing a sample that contains n cases of data into n exploratory sub-samples (each containing n - 1 cases that are obtained from the original n-case sample by the omission of one case without repetition), each of which is used to select a statistical model using the least-squares approach, and n omitted cases, each of which is used to validate the selected model from an exploratory sub-sample that does not contain the omitted case.

An average MSQ is deduced from n models for each subset of candidates. The matrix notation for calculation of MSQ by leave-one-out method is shown in Appendix A. The average MSQs from models of different subsets of candidates are compared, and the model with the smallest average MSQ is taken to be the best subset model. Appendix A shows the matrix notation for the calculation of MSQ by the suggested leave-one-out method.

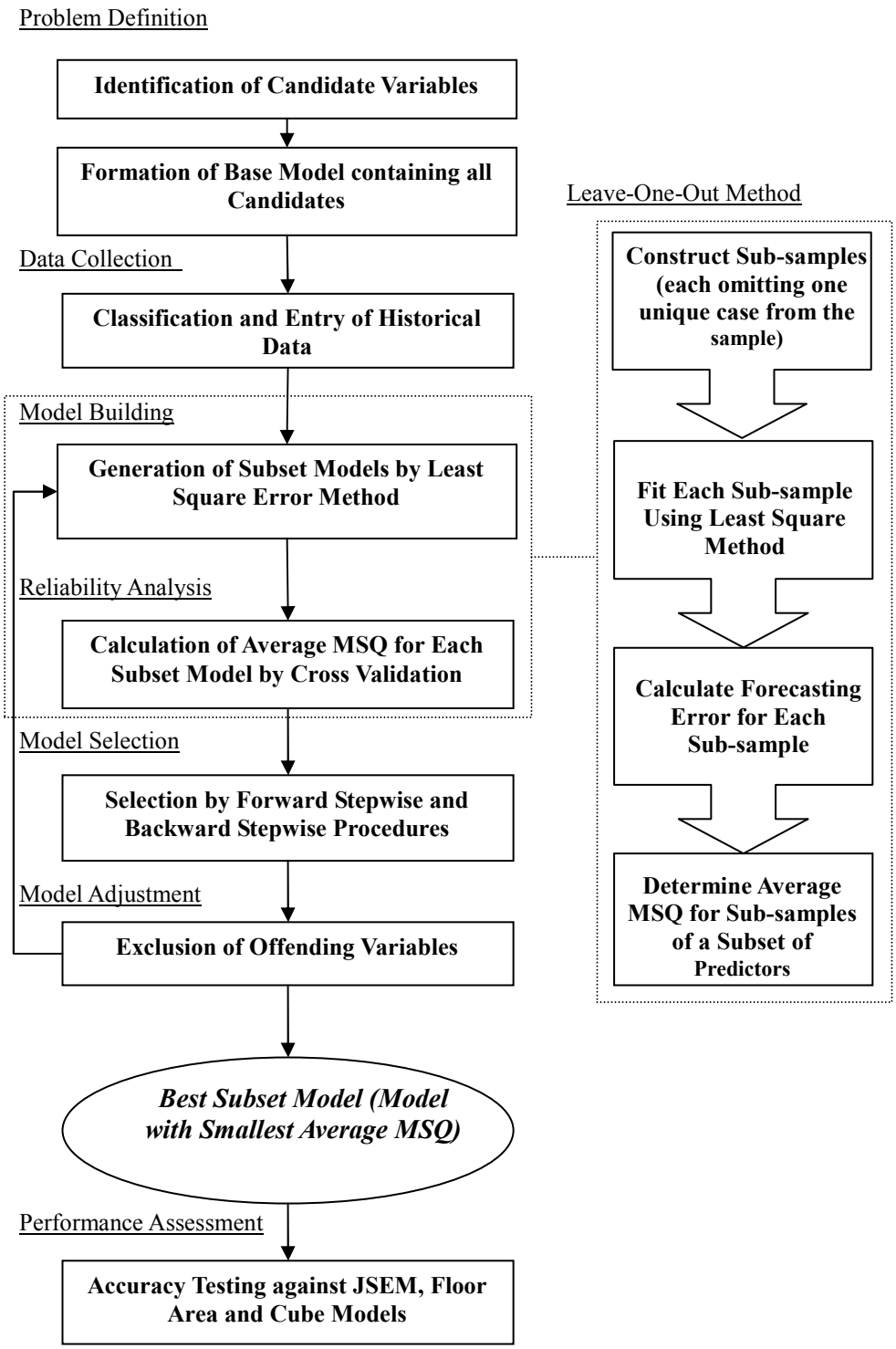

Figure 1: Research Framework for Identification, Selection and Validation of Price Models

7. The Models

To ensure the selection of the best subset model and avoid the time-consuming procedure of trying all the possible combinations, a dual stepwise procedure that merges the forward stepwise and backward stepwise procedures was adopted. According to the algorithm, the selection procedure ends only if the forward and backward stepwise procedures produce the same model (subset of predictors) with the smallest average MSQ. A purpose-made programme written in mathematical software, MathCAD, was used to execute the resampling procedure and the selection algorithm.

Two sets of regression models, namely Regressed Model for James' Storey Enclosure Method (RJSEM) and Regressed Model for Advanced Storey Enclosure Method (RASEM), were developed. The former model uses the various types of enclosure areas in JSEM as candidate variables and the latter model adopts the primary measurements required for deducing the enclosure areas in JSEM as candidate variables. The tender price per total floor area is chosen to be the response of these models because the performance of unit price models can be directly compared with other conventional models on the basis of percentage errors.

Table 2 shows the included candidates, excluded candidates (offending variables excluded by the selection algorithm) and selected predictors for RJSEMs and RASEMs for offices, private housings, nursing homes and schools. A table showing the regression coefficients for each predictor, forecasts and MSQs for the office RASEM is enclosed in Appendix C as an example to illustrate the cross validated results.

8. Performance Validation

A forecast from any of the three conventional models is generated by multiplying the individual quantity to the unit rate deduced by the cross validated method whereas that from the regressed models is automatically generated in a leave-one-out cross validated case. The forecasting error is expressed in percentage terms, i.e. a forecast exceeds the lowest bid, as it is a unit-free measure that is widely adopted as expression of error in practice.

There are two components in interpreting accuracy properly: bias and consistency. Bias is the arithmetic mean of percentage errors, and consistency is measured by the

standard deviation of percentage errors. Table 3 summaries these for the conventional and regressed models for the four types of building.

Since all the models in comparison in this study are generated and tested by the same set of data using cross validation, the forecasted prices generally have very little bias, and most do not deviate significantly from zero. The only exception is the JSEM for offices which is significantly biased at 95% confidence interval, and has the highest mean percentage error amongst all of the models.

To compare models, an algorithm as shown in Figure 2 was used for selecting the appropriate parametric and non-parametric tests.

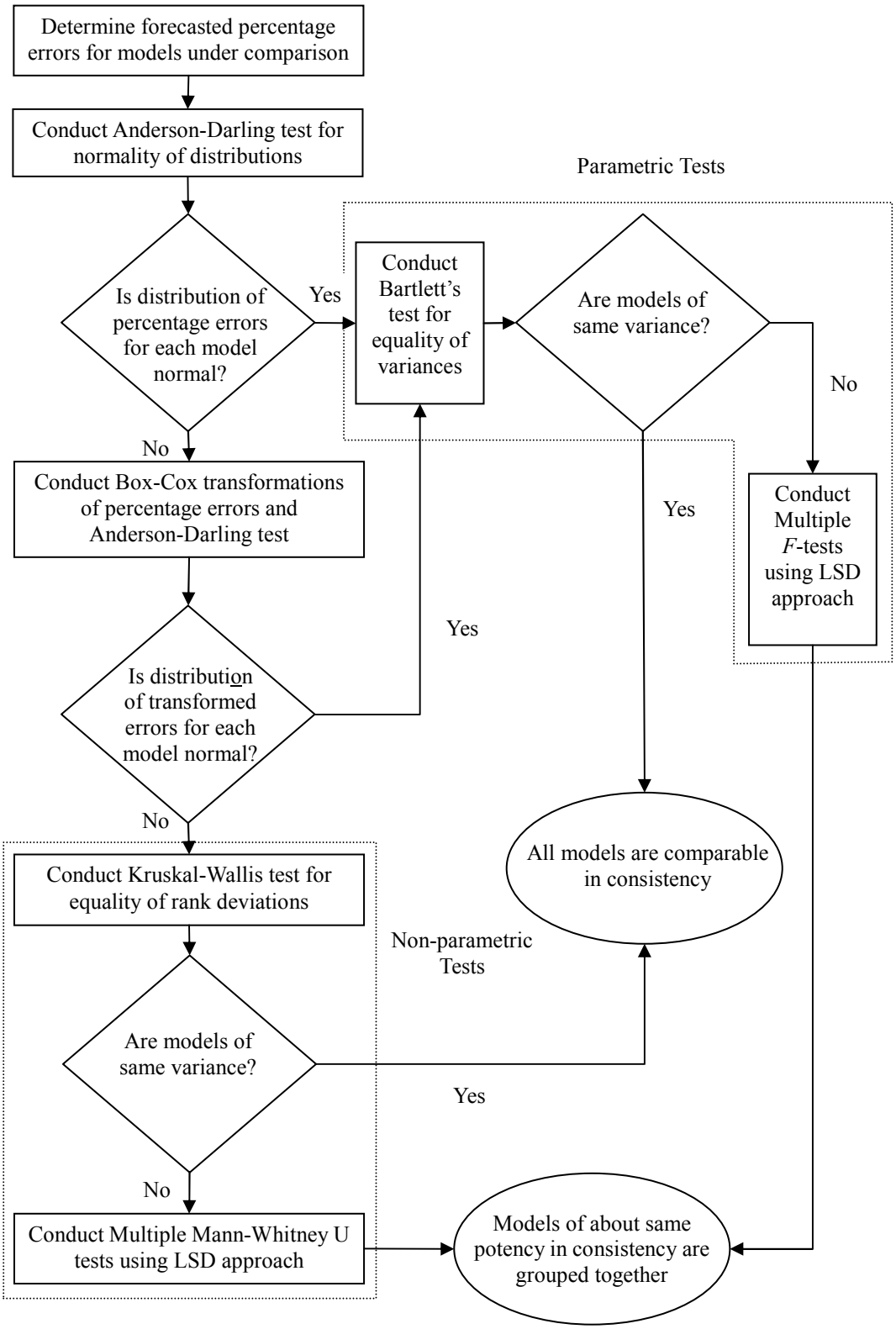

Figure 2: Algorithm for Comparisons of Variances of Percentage Errors

RJSEM and RASEM are grouped individually with the conventional models for comparison. There are eight groups of models: four comprising the RJSEM and the conventional models, and four comprising the RASEM and the conventional models. Each group is first tested for their homogeneity of multivariances by Bartlett's test - the Kruskal-Wallis test (*H*-test) being used as the nonparametric equivalent.

If the variances of models in a group are found to heterogeneous, models are paired up within each group to determine which of the models differ specifically from each other. To do this, the variance of percentage errors of the models are compared in pairwise using the *F*-tests or Mann-Whitney *U* rank sum tests. To correct for any exaggerated significance levels due to multiple testing, Fisher's least significance difference (LSD) approach is used [36].

Figure 3 shows a graphical presentation of the results of these tests. The four groups of models for offices and private housing were found to be significantly different, whereas the four groups for nursing homes and schools were not. Therefore, the former groups were examined in pairwise using Mann-Whitney *U*-tests (see Table 4).

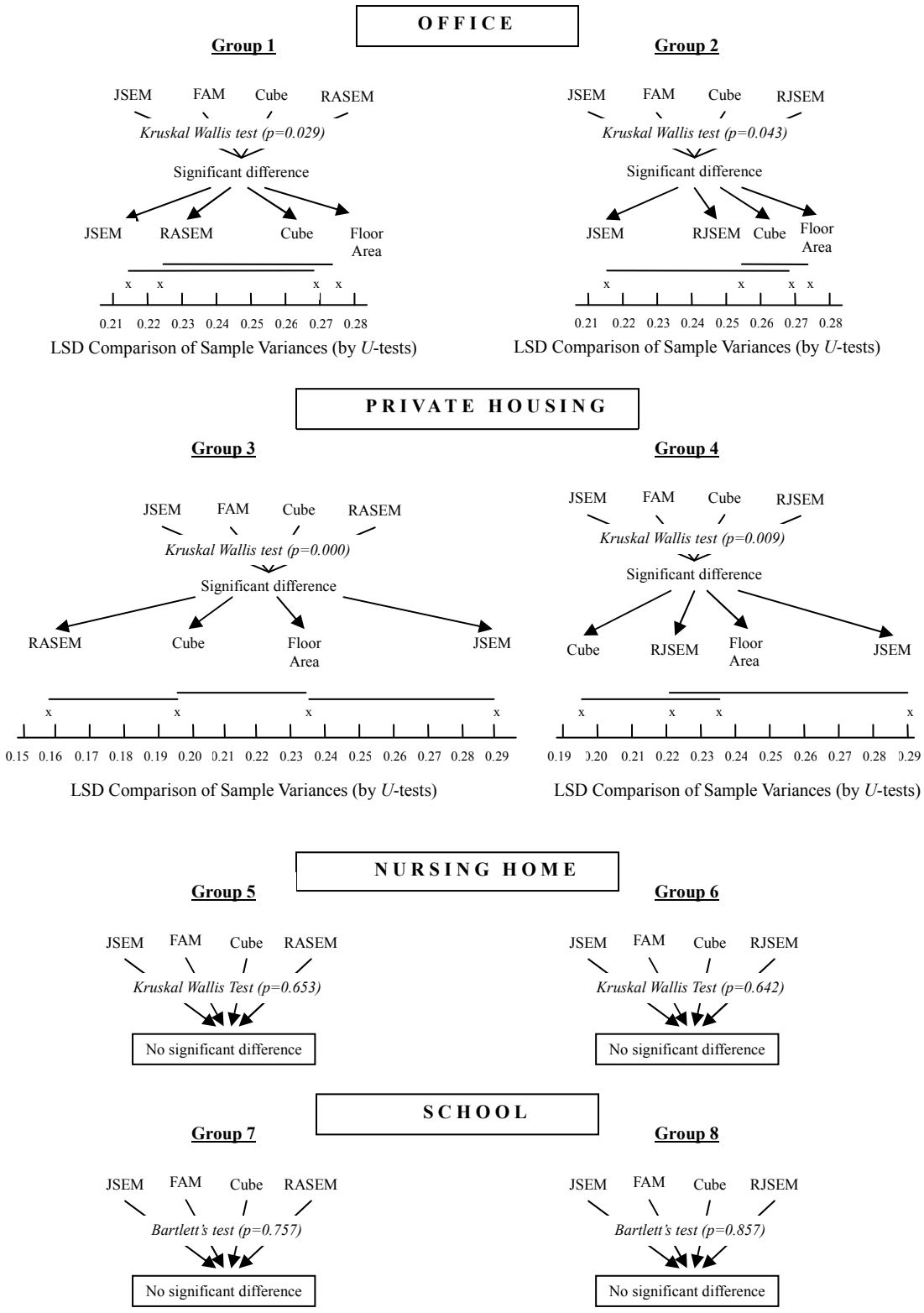

Figure 3: Tests of Homogeneity of Variances Using Bartlett's Tests, Kruskal Wallis Tests and Mann-Whitney U Tests

8.1. Models for Offices (Groups 1 and 2)

Except for the JSEM *t*-tests, all the other *t*-tests for the office models supported the null hypothesis, which suggests that the JSEM is the most biased model and the others are all unbiased models. The Kruskal Wallis tests for both groups of models (as shown in Figure 3) rejected the hypothesis that the models under comparison are equal in consistency. In Group 1, the JSEM, the RASEM and the cube model have the same potency, the RASEM and the cube and FAMs also have the same potency and the JSEM differs from the FAM. Therefore, the more consistent set of models for Group 1 comprises the three comparable models: the JSEM, the RASEM and the cube model. Similarly, the more consistent set of models for Group 2 comprises the JSEM, the RJSEM and the cube model. As JSEM is significantly different from a zero mean percentage error, the best performing sets of models, taking into account both the bias and consistency, are the RASEM and the cube model in Group 1, and the RJSEM and the cube model in Group 2.

8.2. Models for Private Housing (Groups 3 and 4)

All the *t*-tests for the private housing models supported the null hypotheses that the percentage errors of the models are not significantly different from a zero mean. As for Groups 1 and 2, both the Kruskal Wallis tests for the models in Groups 3 and 4 rejected the notion that the models under comparison are equal in consistency. In particular, the RASEM in Group 3 attained relatively spectaculary consistency (15.95%). In this group, the RASEM and the cube model have the same potency, the cube and FAMs have the same potency, the FAM and the JSEM have the same potency, both the RASEM and the cube model differ from the JSEM, and the RASEM differs from the FAM. Therefore, the more consistent set of models for Group 3 comprises the two comparable models: the RASEM and the cube model. In Group 4, the cube model, the RJSEM and the FAM have the same potency, the RJSEM, the FAM and the JSEM have the same potency, and the cube model differs from the JSEM. Therefore, the more consistent set of models for Group 4 comprises the three comparable models: the cube model, RJSEM and FAM.

8.3. Models for Nursing Homes and Schools (Groups 5 to 8)

As with the private housing models, all the *t*-tests for the nursing home and school models supported the null hypotheses that the mean percentage errors of the models are not significantly different from zero. The Kruskal Wallis tests for the models in Groups 5 and 6, and the Bartlett's tests for the models in Groups 7 and 8 both supported the view that the models under comparison are equal in consistency. Therefore, all of the models are comparable with each other in terms of both bias and consistency for Groups 5 to 8. One possible cause for the lack of significant improvement in the regressed models for nursing homes and schools is the insufficient number of candidate variables. In this study, the number of candidates was largely reduced in these two regressed models because of the absence of podiums (for nursing homes and schools) and basements (for schools). Thus, the forecast performance could perhaps be further improved by identifying and including more uncorrelated candidates in the regressed models if more information is extracted as design develops from the early design stage to later stages.

8.4. Discussions on model comparisons

Both the regressed and cube models were included in all of the best sets of comparable models. Rather surprisingly, the popular FAM performs worse than the regressed models, especially for offices. The comparison results, however, created ambiguities in interpreting the models as some models, such as the RASEM and the cube model in Group 1, the RJSEM and the cube model in Group 2, the cube model in Group 3 and the RJSEM and FAM in Group 4, show potency in two different sets of comparable models. Nevertheless, it can be concluded from the LSD comparisons that the use of the RASEM may improve the forecasts, and at least will not worsen them.

The type of information that is available in the early design stage is coarse and very limited, which constricts the forecasting ability of any model, because a model can only capture as much as the available information allows. It appears that even more information such as the elevation area and the roof area, etc. have been extracted and used in the regressed models, the improvement is not significant enough to distinguish them from the conventional models.

9. Conclusion

The cross validation algorithm developed in this study for modelling JSEM's variables makes a significant advancement to the model building process of single rate building price forecasts. Although the data, the observed values for the candidates and the response, used in this study are only for four different types of building projects, the developed methodology for modelling is also applicable to data for other types of buildings as well as other types of data. In using the cross validation approach, both the regressed and conventional models are examined simultaneously based on the same criterion. It is found to be particularly suitable for the problem of building price forecasting, because in practice, forecasters extract the relevant information from a pool of historical projects to make a prediction, and the sample base for modelling in the cross validation approach corresponds to that relevant information. The difference, however, is that practicing forecasters rely heavily on their judgment in choosing the data, the methods for forecasting and deciding the relationship with the tender price. The cross validation approach has considerable intuitive appeal because it produces forecasts in a similar way to forecasters, but it also preserves objectivity.

Conventional approaches to cost modelling generally rely upon the use of historical price data to produce a single-figure (i.e., deterministic) building price forecasts, which do not explicitly describe inherent variability and uncertainty. In the cross validation approach that is used in this research, costs are modelled repetitively, and the reliability of the models is measured according to the mean and standard deviation of percentage errors (the stochastic components of forecasts). The evaluation of the models was conducted with reference to a framework for the selection of the appropriate parametric and non-parametric tests that were used to examine the performance of the models. This framework is an exemplar which ensures the objectivity and rigorousness in the evaluation of models.

Compared with other forecasting regression models, the RASEM and RJSEM gain an advantage over previously developed models in terms of the use of cross validation for reliability analysis, which avoids the major problem of within-sample validation and makes the best use of sample data; applicability, as the candidates and predictors identified are extractable from existing cost analyses, which avoids the subjective elements in defining and measuring qualitative variables; and the use of statistical inference for comparing models, which provides a fair basis for the assessment of model performance. Although the regressed models are not distinguishably better, they are replicable, because they are backed by the cross validation approach; are easy to use,

because they involve only a few predictors; and are fairly accurate and reliable, because they are comparable with other models within the best clusters. If a cross validated regression model is chosen for prediction, then it can, on average, produce forecasts that are at least as good as the forecasts of any of the conventional models. For certain applications, such as when the RASEM is used for private housing, the chance of getting better forecasts is high.

Once a cross validated model is developed, both experienced and inexperienced forecasters should be able to apply the models without any difficulties. Practitioners may also refer to the methodology described in this paper to develop cost models by inputting different set of variables such as the number of bidders and type of contract to satisfy the forecasting needs in the later stages.

Appendix A: Matrix Notation for Calculation of MSQ by Leave-one-out Method

Let $\boldsymbol{P}$ be a column vector containing n rows of observed values for the response $\{P_1, P_2, \ldots, P_n\}^T$ and $\underline{\boldsymbol{V}}$ be a matrix that contains n x $(k + 1)$ of the observed values for a subset of variables such that:

$$\underline{\boldsymbol{V}} = \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \vdots \\ \boldsymbol{V}_n \end{bmatrix} = \begin{bmatrix} 1 & V_{1,1} & V_{1,2} & \cdots & V_{1,k} \\ 1 & V_{2,1} & V_{2,2} & \cdots & V_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & V_{n,1} & V_{n,2} & \cdots & V_{n,k} \end{bmatrix}. \qquad \text{(A.1)}$$

Corresponding to P_i is $\boldsymbol{V}_i$, a row vector that contains the observed values for the variables (which contain a constant term and k number of predictors) and $\{1, V_{i,1}, V_{i,2}, \ldots, V_{i,k}\}$, where $i = 1, 2, \ldots, n$. In a regressed model, the price is represented by:

$$\boldsymbol{P} = \underline{\boldsymbol{V}}\boldsymbol{\beta} + \boldsymbol{e}, \qquad \text{(A.2)}$$

where $\boldsymbol{\beta}$ is a column vector of the coefficients $\{\beta_0, \beta_1, \beta_2, \ldots, \beta_k\}^T$ and $\boldsymbol{e}$ is a column vector of the forecasting errors $\{e_1, e_2, \ldots, e_n\}^T$. The mean square error then becomes:

$$\left.\begin{aligned} \text{MSQ} &= \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}(\boldsymbol{e}^T\boldsymbol{e}) \\ &= \frac{1}{n}\left(\boldsymbol{P} - \underline{\boldsymbol{V}}\boldsymbol{\beta}\right)^T\left(\boldsymbol{P} - \underline{\boldsymbol{V}}\boldsymbol{\beta}\right) \\ &= \frac{1}{n}(\boldsymbol{P}^T\boldsymbol{P} - \boldsymbol{\beta}^T\underline{\boldsymbol{V}}^T\boldsymbol{P} - \boldsymbol{P}^T\underline{\boldsymbol{V}}\boldsymbol{\beta} + \boldsymbol{\beta}^T\underline{\boldsymbol{V}}^T\underline{\boldsymbol{V}}\boldsymbol{\beta}) \end{aligned}\right\} \qquad \text{(A.3)}$$

$\hat{\boldsymbol{\beta}}$ is the $\boldsymbol{\beta}$ that produces the minimum MSQ. To determine $\hat{\boldsymbol{\beta}}$, the MSQ is differentiated with respect to $\boldsymbol{\beta}$, and the result is equated to zero, i.e.,

$$\left.\frac{\partial \mathrm{MSQ}}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \frac{1}{n}(-2\underline{\boldsymbol{V}}^T\boldsymbol{P} + 2\underline{\boldsymbol{V}}^T\underline{\boldsymbol{V}}\hat{\boldsymbol{\beta}}) = \boldsymbol{0}. \qquad \text{(A.4)}$$

This yields:

$$\left.\begin{aligned} &\underline{\boldsymbol{V}}^T\underline{\boldsymbol{V}}\hat{\boldsymbol{\beta}} = \underline{\boldsymbol{V}}^T\boldsymbol{P} \\ &\hat{\boldsymbol{\beta}} = \left(\underline{\boldsymbol{V}}^T\underline{\boldsymbol{V}}\right)^{-1}\underline{\boldsymbol{V}}^T\boldsymbol{P} \end{aligned}\right\} \qquad \text{(A.5)}$$

Therefore, the minimum MSQ is:

$$\mathrm{MSQ}_{min} = \frac{1}{n}\left(\boldsymbol{P}^T\boldsymbol{P} - \hat{\boldsymbol{\beta}}^T\underline{\boldsymbol{V}}^T\boldsymbol{P} - \boldsymbol{P}^T\underline{\boldsymbol{V}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\beta}}^T\underline{\boldsymbol{V}}^T\underline{\boldsymbol{V}}\hat{\boldsymbol{\beta}}\right). \qquad \text{(A.6)}$$

Referring to the least-squares method that is described in the matrix notation above, let $\boldsymbol{P}^{(-j)}$ be a column vector that contains n rows of observed values for the response $\{P_1, P_2, \ldots, P_{(j-1)}, P_{(j+1)}, \ldots, P_n\}^T$, let $\underline{\boldsymbol{V}}^{(-j)}$ be a matrix containing $(n-1)$ x $(k+1)$ of the observed values for the subset of variables (with the omission of one row of the observed values, representing the j^{th} case, from the matrix of variables $\underline{\boldsymbol{V}}$ such that j is any number from 1 to n):

$$\boldsymbol{V}^{(-j)} = \begin{bmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \vdots \\ \boldsymbol{V}_{(j-1)} \\ \boldsymbol{V}_{(j+1)} \\ \vdots \\ \boldsymbol{V}_n \end{bmatrix} = \begin{bmatrix} 1 & V_{1,1} & V_{1,2} & \cdots & V_{1,k} \\ 1 & V_{2,1} & V_{2,2} & \cdots & V_{2,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & V_{(j-1),1} & V_{(j-1),2} & \cdots & V_{(j-1),k} \\ 1 & V_{(j+1),1} & V_{(j+1),2} & \cdots & V_{(j+1),k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & V_{n,1} & V_{n,2} & \cdots & V_{n,k} \end{bmatrix}. \qquad \text{(A.7)}$$

$\boldsymbol{\beta}^{(-j)}$ is a column vector of the coefficients $\{\beta_0, \beta_1, \beta_2, \ldots, \beta_{(j-1)}, \beta_{(j+1)}, \ldots, \beta_k\}^T$ and $\boldsymbol{e}^{(-j)}$ is a column vector of the forecasting errors $\{e_1, e_2, \ldots, e_{(j-1)}, e_{(j+1)}, \ldots, e_n\}^T$ of the regressed model $\boldsymbol{P}^{(-j)} = \boldsymbol{V}^{(-j)}\boldsymbol{\beta}^{(-j)} + \boldsymbol{e}^{(-j)}$. Similar to the derivation that is shown in

Equations (A.3) to (A.6), the minimum MSQ of the regressed model that does not contain the j^{th} case becomes:

$$\mathrm{MSQ}_{min}^{(-j)} = \frac{1}{n}\left(\boldsymbol{P}^{(-j)\boldsymbol{T}}\boldsymbol{P}^{(-j)} - \hat{\boldsymbol{\beta}}^{(-j)\boldsymbol{T}}\underline{\boldsymbol{V}}^{(-j)\boldsymbol{T}}\boldsymbol{P}^{(-j)} - \boldsymbol{P}^{(-j)\boldsymbol{T}}\underline{\boldsymbol{V}}^{(-j)}\hat{\boldsymbol{\beta}}^{(-j)} + \hat{\boldsymbol{\beta}}^{(-j)\boldsymbol{T}}\underline{\boldsymbol{V}}^{(-j)\boldsymbol{T}}\underline{\boldsymbol{V}}^{(-j)}\hat{\boldsymbol{\beta}}^{(-j)} \right) \quad \text{(A.8)}$$

The average of $\mathrm{MSQ}_{min}^{(-j)}$, $\overline{\mathrm{MSQ}_{min}^{(-j)}}$, is deduced from n regressed models (for $j = 1, \ldots, n$) of the subset of variables in accordance with Equation (A.9),

$$\overline{\mathrm{MSQ}_{min}^{(-j)}} = \frac{1}{n}\sum_{j=1}^{n}\mathrm{MSQ}_{min}^{(-j)} . \quad \text{(A.9)}$$

Different $\overline{\mathrm{MSQ}_{min}^{(-j)}}$ from different subsets of variables that are chosen by the selection strategy that is described in the next section are compared. The subset of variables that gives the smallest $\overline{\mathrm{MSQ}_{min}^{(-j)}}$ is the best subset model.

Appendix B: Original Data

Table B-1: Original Data for Offices

Case	No. of floor for pdm.	No. of floor for tower	No. of floor for bmt.	Avg. area per floor for pdm.	Avg. area per floor for tower	Avg. area per floor for bmt.	Avg. storey height of pdm.	Avg. storey height of tower	Avg. storey height of bmt.	Avg. perim. on plan for pdm. and tower	Avg. perim. on plan for bmt.	Roof area	Adjusted tender price*
	(a)	*(b)*	*(m)*	*(fp)*	*(ft)*	*(fb)*	*(sp)*	*(st)*	*(sb)*	*(ppt)*	*(pb)*	*(r)*	*(TP)*
1	5	14	1	696	568	785	3	3	6	129	130	1150	5.72E+07
2	4	17	1	580	373	518	4	3	5	93	110	866	4.33E+07
3	5	12	1	1500	1180	1670	3	3	5	146	158	2430	9.85E+07
4	3	21	0	556	420	0	4	3	0	92	0	887	4.97E+07
5	4	22	0	669	452	0	5	3	0	105	0	1120	5.88E+07
6	0	13	1	0	3600	4140	0	3	4	245	275	3600	2.08E+08
7	4	22	1	1230	885	662	4	4	5	143	110	2120	1.17E+08
8	4	20	0	400	246	0	5	4	0	103	0	698	5.96E+07
9	0	70	2	0	1630	7980	0	5	8	215	545	1630	1.05E+09
10	2	23	3	1370	1440	2130	5	4	7	229	181	2800	3.68E+08
11	3	21	5	1100	1520	2520	4	4	6	195	168	2620	3.23E+08
12	3	26	1	341	329	434	4	4	5	87	104	676	7.20E+07
13	5	29	0	569	257	0	4	3	0	92	0	774	5.75E+07
14	3	32	2	12800	7380	10400	5	3	5	596	619	20200	1.48E+09
15	3	22	1	1180	895	1370	5	4	5	194	200	2070	2.25E+08
16	4	15	1	410	240	340	4	4	4	88	97	619	4.75E+07
17	5	17	1	4120	3110	8230	5	4	7	246	257	7230	4.08E+08
18	4	21	0	536	302	0	4	3	0	80	0	798	3.95E+07
19	0	23	0	0	160	0	0	3	0	62	0	160	2.88E+07
20	4	34	0	907	489	0	7	4	0	146	0	1400	1.78E+08
21	0	21	0	0	261	0	0	3	0	71	0	249	2.50E+07
22	6	18	0	194	148	0	3	3	0	54	0	342	2.43E+07
23	3	23	1	1200	928	1510	6	4	5	218	157	2130	2.38E+08
24	5	19	0	503	326	0	4	3	0	78	0	802	4.72E+07
25	3	34	2	4110	3250	3620	4	3	6	301	303	7360	5.65E+08
26	4	24	1	481	339	1500	4	4	5	109	155	1090	1.10E+08
27	5	20	1	307	286	395	3	3	4	74	98	705	5.60E+07
28	2	24	0	757	411	0	4	4	0	104	0	1170	7.83E+07
29	4	18	0	1240	558	0	3	3	0	124	0	1710	8.79E+07
30	3	20	0	3140	1370	0	3	3	0	177	0	4510	1.62E+08
31	0	9	0	0	950	0	0	4	0	105	0	760	3.50E+07
32	5	19	0	347	315	0	4	4	0	125	0	698	4.72E+07
33	4	13	1	334	250	335	4	3	3	61	74	557	2.66E+07
34	4	14	1	482	291	454	4	3	5	83	88	736	3.80E+07
35	5	12	0	325	304	0	4	3	0	77	0	571	2.44E+07
36	3	41	4	3884	3450	7590	5	4	4	326	510	7560	1.30E+09
37	3	69	3	2900	2060	7040	6	4	4	224	336	4960	9.39E+08
38	0	14	0	0	1060	0	0	4	0	141	0	1060	6.22E+07
39	2	8	0	1730	601	0	4	5	0	105	0	2330	3.92E+07
40	0	17	0	0	890	0	0	4	0	132	0	890	6.23E+07
41	0	7	1	0	4650	2800	0	5	5	385	215	4650	1.53E+08
42	0	25	2	0	616	749	0	4	4	106	125	616	8.09E+07

Remarks:

* - Adjusted tender prices are rebased to the price level in the 2nd quarter of 1997 with reference to (Tender Price Indices and Cost Trends produced by Levett and Bailey Chatered Quantity Surveyors Ltd.

Table B-2: Original Data for Private Housing

Case	No. of floor for pdm.	No. of floor for tower	No. of floor for bmt.	Avg. area per floor for pdm.	Avg. area per floor for tower	Avg. area per floor for bmt.	Avg. storey height of pdm.	Avg. storey height of tower	Avg. storey height of bmt.	Avg. perim. on plan for pdm. and	Avg. perim. on plan for bmt.	Roof area	Adjusted tender price*
	(a)	*(b)*	*(m)*	*(fp)*	*(ft)*	*(fb)*	*(sp)*	*(st)*	*(sb)*	*(ppt)*	*(pb)*	*(r)*	*(TP)*
1	1	39	0	4960	2920	0	5	3	0	1110	0	4960	6.52E+08
2	1	41	0	7960	3590	0	4	3	0	1370	0	7960	9.26E+08
3	3	21	0	3070	1030	0	4	3	0	319	0	3070	1.77E+08
4	7	32	0	1010	433	0	3	4	0	144	0	1010	1.23E+08
5	1	33	0	2300	5440	0	4	3	0	1420	0	2300	6.15E+08
6	4	52	2	24900	2170	14000	4	3	3	564	2440	24900	1.11E+09
7	2	14	0	350	150	0	4	3	0	78	0	350	1.57E+07
8	3	13	0	696	306	0	3	3	0	123	0	696	3.03E+07
9	3	44	0	3110	765	0	4	3	0	272	0	3110	1.88E+08
10	4	28	1	15000	2290	10100	3	3	3	990	1450	15000	7.36E+08
11	0	37	1	0	5320	14600	0	3	3	1410	2120	5320	1.29E+09
12	0	33	1	0	3690	10300	0	3	4	983	1580	3690	5.40E+08
13	6	24	0	2390	833	0	3	3	0	235	0	2390	2.02E+08
14	4	29	0	346	131	0	3	3	0	66	0	346	2.96E+07
15	6	20	0	489	244	0	3	3	0	85	0	489	3.70E+07
16	3	38	0	10800	4340	0	4	3	0	1690	0	10800	6.18E+08
17	4	11	0	338	131	0	3	3	0	62	0	338	1.14E+07
18	3	38	0	5300	3300	0	3	3	0	826	0	5300	5.30E+08
19	2	16	0	910	314	0	3	3	0	127	0	910	2.46E+07
20	0	21	1	0	1360	2960	0	3	3	469	544	1360	1.71E+08
21	3	35	1	2270	1690	865	4	3	3	557	168	2270	3.35E+08
22	0	34	0	0	1350	0	0	3	0	484	0	1350	2.19E+08
23	2	33	2	2310	1540	2350	3	3	3	461	551	2310	1.90E+08
24	2	36	0	12500	4510	0	4	3	0	1330	0	12500	7.74E+08
25	4	20	0	328	123	0	2	3	0	61	0	328	1.73E+07
26	2	15	1	7350	5100	4570	3	3	4	1580	799	7350	3.56E+08
27	0	16	0	0	1290	0	0	3	0	381	0	1290	8.31E+07
28	4	37	0	3810	2040	0	3	3	0	745	0	3810	2.93E+08
29	1	23	1	4340	4500	787	4	3	3	1340	138	4340	3.91E+08
30	2	18	1	2130	1560	2130	4	3	3	469	445	2130	1.86E+08
31	0	32	1	0	3760	2430	0	3	4	1090	303	3760	4.14E+08
32	0	31	0	0	1050	0	0	3	0	373	0	1050	9.89E+07
33	3	36	0	8700	1350	0	3	3	0	678	0	8700	3.72E+08
34	3	30	0	5420	2100	0	3	3	0	830	0	5420	2.78E+08
35	3	36	0	5500	2260	0	3	3	0	706	0	5500	3.86E+08
36	2	29	4	932	488	1480	4	3	3	175	303	932	8.99E+07
37	0	37	0	0	3350	0	0	3	0	708	0	3350	3.80E+08
38	0	39	0	0	4340	0	0	3	0	1270	0	4340	5.11E+08
39	0	37	0	0	2320	0	0	3	0	941	0	2320	2.63E+08
40	0	37	0	0	4560	0	0	3	0	1690	0	4560	5.14E+08
41	0	37	0	0	3470	0	0	3	0	1270	0	3470	4.23E+08
42	0	36	0	0	2430	0	0	3	0	904	0	2430	3.69E+08
43	0	37	0	0	3510	0	0	3	0	1540	0	3510	4.53E+08
44	0	35	0	0	603	0	0	3	0	196	0	603	7.54E+07
45	0	37	0	0	1910	0	0	3	0	611	0	1910	2.19E+08
46	1	24	0	1600	1100	0	0	3	0	371	0	1600	7.89E+07
47	0	26	0	0	956	0	0	3	0	303	0	956	1.29E+08
48	0	19	1	0	754	610	0	3	3	253	118	754	6.41E+07
49	0	14	0	0	1450	0	0	3	0	530	0	1450	7.83E+07
50	0	39	0	0	3300	0	0	3	0	1030	0	3300	3.90E+08

Remarks:

* - Adjusted tender prices are rebased to the price level in the 2nd quarter of 1997 with reference to (Tender Price Indices and Cost Trends produced by Levett and Bailey Chatered Quantity Surveyors Ltd.

Table B-3: Original Data for Nursing Home

Case	No. of floor for pdm. and tower	No. of floor for btm.	Avg. area per floor for pdm. and tower	Avg. area per floor for btm.	Avg. storey height of pdm. and tower	Avg. storey height of btm.	Avg. perim. on plan for pdm. and tower	Avg. perim. on plan for btm.	Roof area	Adjusted tender price*
	(n)	*(m)*	*(fpt)*	*(fb)*	*(spt)*	*(sb)*	*(ppt)*	*(pb)*	*(r)*	*(TP)*
1	8	1	1150	157	4	3	218	53	1250	4.24E+07
2	9	1	1040	1280	4	3	210	356	1040	3.63E+07
3	6	0	1490	0	4	0	254	0	1590	4.49E+07
4	5	0	2020	0	5	0	249	0	2020	4.26E+07
5	6	0	400	0	4	0	78	0	460	1.25E+07
6	8	0	655	0	4	0	164	0	655	2.25E+07
7	7	0	500	0	4	0	290	0	550	2.06E+07
8	6	0	631	0	4	0	137	0	631	2.09E+07
9	6	1	1590	1360	3	3	254	360	1590	4.24E+07
10	8	0	858	0	4	0	167	0	878	2.73E+07
11	5	0	1230	0	3	0	275	0	1430	2.86E+07
12	5	0	2380	0	4	0	339	0	2580	4.03E+07
13	5	0	915	0	4	0	248	0	915	1.86E+07
14	4	0	1430	0	3	0	342	0	1530	1.61E+07
15	6	0	959	0	4	0	125	0	959	1.83E+07
16	4	0	935	0	5	0	155	0	935	1.73E+07
17	24	2	594	860	3	3	155	273	594	8.56E+07
18	3	0	2080	0	3	0	439	0	2480	1.82E+07
19	11	0	555	0	3	0	131	0	555	2.60E+07
20	7	0	629	0	4	0	114	0	639	2.23E+07
21	5	0	858	0	4	0	156	0	858	1.34E+07
22	6	0	872	0	3	0	252	0	872	1.70E+07
23	11	1	553	1110	3	4	190	257	553	3.90E+07

Remarks:
* - Adjusted tender prices are rebased to the price level in the 2nd quarter of 1997 with reference to (Tender Price Indices and Cost Trends produced by Levett and Bailey Chatered Quantity Surveyors Ltd.

Table B-4: Original Data for School

Case	No. of floor for pdm. and tower	Avg. area per floor for pdm. and tower	Avg. storey height of pdm. and tower	Avg. perimeter on plan for pdm. and tower	Roof area	Adjusted tender price*
	(n)	*(fpt)*	*(spt)*	*(ppt)*	*(r)*	*(TP)*
1	5	900	4	183	900	1.03E+07
2	5	2493	4	350	2293	2.39E+07
3	5	900	4	183	900	1.10E+07
4	5	887	4	220	787	1.06E+07
5	3	903	4	162	903	5.07E+06
6	6	495	4	108	495	6.67E+06
7	5	554	4	130	554	5.26E+06
8	6	1619	4	370	1619	1.99E+07
9	7	796	4	180	796	1.87E+07
10	4	870	5	207	870	7.06E+06
11	6	511	4	124	511	7.78E+06
12	5	963	4	220	963	9.34E+06
13	5	892	4	176	892	1.46E+07
14	4	1085	4	289	1285	1.02E+07
15	5	1665	4	354	1665	1.57E+07
16	7	1024	3	194	1024	1.42E+07
17	2	2100	4	477	2100	6.86E+06
18	3	1051	4	255	1051	5.87E+06
19	4	535	4	110	535	3.23E+06
20	4	980	4	277	980	9.38E+06
21	4	494	4	108	494	4.19E+06
22	3	974	3	247	974	4.63E+06
23	4	2425	4	513	2225	1.48E+07

Remarks:
* - Adjusted tender prices are rebased to the price level in the 2nd quarter of 1997 with reference to (Tender Price Indices and Cost Trends produced by Levett and Bailey Chatered Quantity Surveyors Ltd.

Appendix C: Coefficients, Forecasts and MSQs for RJSEMs and RASEM

Table C-1: Coefficients, Forecasts and MSQs Determined by Leave-One-Out Method for the RASEM for Offices

Case	$RASEM\ (\beta_0 + \beta_1 \cdot nspt + \beta_2 \cdot n2 + \beta_3 \cdot fpt + \beta_4 \cdot ppt)$						
	β_0	β_1	β_2	β_3	β_4	Forecasted Y	MSQ
1	2370	43.45	-1.892	-1.571	18.76	6,005	1.75E+06
2	2359	45.68	-1.981	-1.450	16.88	5,576	7.34E+05
3	2363	45.18	-1.961	-1.454	17.08	4,937	5.14E+05
4	2356	45.81	-1.983	-1.441	16.76	5,520	6.09E+05
5	2300	48.45	-2.107	-1.398	15.94	6,487	3.34E+06
6	2025	47.18	-2.038	-1.688	19.90	2,394	2.84E+06
7	2288	47.24	-2.057	-1.453	16.94	6,433	3.11E+06
8	2395	42.12	-1.799	-1.468	17.60	7,068	4.31E+06
9	2235	47.75	-2.025	-1.407	16.24	8,550	2.40E+05
10	2405	48.12	-2.059	-1.218	13.33	7,416	1.68E+06
11	2281	46.65	-2.024	-1.440	16.83	6,899	2.23E+04
12	2286	45.11	-1.955	-1.480	17.57	6,409	6.10E+05
13	2298	47.77	-2.061	-1.402	16.05	6,610	1.06E+06
14	2387	46.10	-1.982	-1.508	16.59	3,836	1.36E+06
15	2433	46.52	-1.989	-1.260	14.10	7,866	1.65E+06
16	2106	48.58	-2.091	-1.368	16.09	5,694	7.93E+06
17	2290	46.27	-2.004	-1.449	16.94	4,925	4.25E+03
18	2361	46.88	-2.029	-1.396	15.99	5,738	1.18E+06
19	2096	45.67	-1.983	-1.546	18.86	5,414	5.84E+06
20	2239	47.86	-2.076	-1.422	16.50	9,053	6.92E+04
21	2370	46.51	-2.016	-1.405	16.15	5,519	9.40E+05
22	2218	45.53	-1.976	-1.510	18.07	5,430	8.50E+05
23	2352	46.75	-2.015	-1.353	15.49	8,587	1.70E+05
24	2300	46.61	-2.019	-1.427	16.61	5,587	2.48E+04
25	2291	45.53	-1.963	-1.448	17.38	5,605	1.60E+06
26	2269	46.35	-2.005	-1.340	15.72	6,509	9.02E+06
27	2108	47.33	-2.061	-1.478	17.69	5,339	3.96E+06
28	2292	46.35	-2.008	-1.438	16.80	6,770	1.22E+04
29	2275	46.94	-2.034	-1.420	16.53	5,661	3.95E+04
30	2335	46.00	-2.001	-1.440	16.93	5,606	1.46E+06
31	2293	46.50	-2.015	-1.434	16.73	4,113	5.50E+02
32	2296	45.41	-1.973	-1.516	17.92	6,801	4.78E+05
33	2266	46.47	-2.013	-1.448	16.99	5,171	5.72E+04
34	2244	47.00	-2.035	-1.430	16.74	5,392	2.48E+05
35	2347	45.84	-1.987	-1.440	16.76	5,121	2.41E+05
36	2270	46.98	-2.035	-1.425	16.63	7,224	1.80E+04
37	2101	52.95	-2.439	-1.337	15.23	4,320	1.30E+06
38	2367	45.35	-1.982	-1.469	17.29	5,611	2.02E+06
39	2284	46.62	-2.020	-1.434	16.73	4,696	1.96E+03
40	2378	44.97	-1.963	-1.488	17.53	5,720	2.56E+06
41	2290	46.48	-2.015	-1.435	16.77	4,363	1.06E+03
42	2202	53.66	-2.347	-1.232	13.35	6,966	4.73E+06
Average:							**1.63E+06**

References

[1] James W. A New Approach to Single Price Rate Approximate Estimating. RICS Journal XXXIII (XI) 1954;. (May): 810-24.

[2] Department of Environment (DOE) Local Authority Offices: Areas and Costs. London: DOE; 1971.

[3] Tregenza T. Association between building height and cost. Architects' Journal 1972; 156(44): 1031-2.

[4] Flanagan R, Norman G. The relationship between construction price and height. Chartered Surveyor B and QS Quarterly 1978; 69-71.

[5] Karshenas S. Predesign Cost Estimating Method for Multi-storey Buildings. Journal of Construction Engineering and Management, ASCE 1984; 110(1): 79-86.

[6] Skitmore RM, Patchell BRT. Development in contract price forecasting and bidding techniques. In: Brandon P.S. editor. Quantity Surveying Techniques: New Directions. Blackwell Scientific; 1990, p. 75-120.

[7] Moyles BF. An Analysis of the Contractors' Estimating Process. Department of Civil Engineering, Loughborough University of Technology; 1973.

[8] Neale RH. The Use of Regression Analysis as a Contractor's Estimating Tool. Department of Civil Engineering, Loughborough University of Technology; 1973.

[9] Braby RH. Costs of high-rise buildings. Building Economist 1975; 14: 84-86.

[10] Khosrowshahi F, Kaka AP. Estimation of Project Total Cost and Duration for Housing Projects in the U.K. Building and Environment 1996; 31(4): 375-383.

[11] Baker MJ. Cost of Houses for the Aged. Department of Civil Engineering, Loughborough University of Technology; 1974.

[12] Blackhall JD. The Application of Regression Modelling to the Production of a Price Index for Electrical Services. Department of Civil Engineering, Loughborough University of Technology; 1974.

[13] Coates D. Estimating for French Drains - A Computer Based Model. Department of Civil Engineering, Loughborough University of

Technology; 1974.

[14] Kouskoulas V, Koehn E. Predesign cost estimating function for buildings. Journal of Construction Division, ASCE 1974; 589-604.

[15] Buchanan JS. Cost Models for Estimating: Outline of the Development of a Cost Model for the Reinforced Concrete Frame of a Building. London, RICS; 1972.

[16] Singh S. Computer Model for Cost Estimation of Structures in High Rise Commercial Buildings. Proceedings of the Annual Conference – Associate Schools of Construction 1989 25th; 1989.

[17] Gould PR. The Development of a Cost Model for H&V and A. C. Installations in Buildings. Department of Civil Engineering, Loughborough University of Technology; 1970.

[18] Southwell J. Building Cost Forecasting, Royal Institution of Chartered Surveyors; 1971.

[19] Mathur K. editor. A Probabilistic Planning Model. Building cost techniques: New Directions, E & FN Spon; 1982.

[20] Pitt T. The Identification and Use of Spend Units in the Financial Monitoring and Control of Construction Projects. In: Brandon P.S. editor. Building Cost Techniques - New Directions. London: E & F N Spon; 1982, p. 255-62.

[21] Wilson AJ. Experiments in probabilistic cost modelling. In: Brandon P.S. editor. Building Cost Techniques - New Directions. London: E & F N Spon; 1982, p. 169-80.

[22] Bennett J, Omerod RN. Simulation Applied to Construction Projects. Construction Management and Economics 1984; 2: 225-63.

[23] Chau KW. Monte Carlo simulation of construction costs using subjective data. Construction Management and Economics 1995; 13: 369-83.

[24] Chau KW. The validity of the triangular distribution assumption in Monte Carlo simulation of construction costs: empirical evidence from Hong Kong. Construction Management and Economics 1995; 13: 15-21.

[25] Li H. Neural networks for construction cost estimation. Building Research and Information 1995; 23(5): 279-284.

[26] Adeli H, Wu M. Regularization neural network for construction cost estimation. Journal of Construction Engineering and Management 1998;

24(1): 18-24.

[27] Bode J. Neural networks for cost estimation. Cost Engineering 1998; 40(1): 25-30.

[28] Emsley MW, Lowe DJ, Duff AR, Harding A, Hickson A. Data modelling and the application of a neural network approach to the prediction of total construction costs. Construction Management and Economics 2002; 20(6): 465-472.

[29] Kim G.H, Yoon JE, An SH, Cho HH, Kang KI. Neural network model incorporating a genetic algorithm in estimating construction costs. Building and Environment 2004; 39(11): 1330-1340.

[30] Lu Q. Cost estimation based on theory of fuzzy sets and prediction techniques - an expert system approach. Construction Contracting in China, Department of Civil and Structural Engineering, Hong Kong Polytechnic; 1988, p.113-25.

[31] Fortune C, Lees M. The relative performance of new and traditional cost models in strategic advice and clients, The Royal Institution of Chartered Surveyors; 1996.

[32] Fortune C, Hinks J. Strategic building project price forecasting models in use - paradigm shift postponed. Journal of Financial Management of Property and Construction 1998; 3(1): 3-26.

[33] Bowen PA, Edwards PJ. Building Cost Planning and Cost Information Management in South Africa. International Journal of Procurement 1998; (June): 16-25.

[34] Bowen PA, Edwards PJ. Cost Modelling and Price Forecasting: Practice and Theory in Perspective. Construction Management and Economics 1985; 3: 199-215.

[35] Raftery J. The state of cost/price modelling in the construction industry: a multicriteria approach. In: Brandon P.S. editor. Building Cost Modelling and Computers. E & F N Spon; 1987, p. 49-71.

[36] Kleinbaum DG., Kupper LL, Muller KE. Applied regression analysis and other multivariable methods. 3rd ed. Boston, Mass.:, PWS-Kent; 1998, p. 43-46, p. 443-447

[37] McLachlan GJ. Error rate estimation in discriminant analysis: Recent advances. In: Gupta A.K. editor. Advances in Multivariate Statistical

Analysis. Dordrect, The Netherlands: Reidel; 1987, p. 233-252.

[38] Skitmore M. Parameter prediction for cash flow forecasting models. Construction Management and Economics 1992; 10: 397-413.

[39] Goutte C. Note on free lunches and cross-validation. Neural Computation 1997; 9(6): 1246-9.

Figure Captions

Figure 1: Research Framework for Identification, Selection and Validation of Price Models

Figure 2: Algorithm for Comparisons of Variances of Percentage Errors

Figure 3: Tests of Homogeneity of Variances Using Bartlett's Tests, Kruskal Wallis Tests and Mann-Whitney U Tests

Tables

Table 1: List of Candidate Variables

Primary Model	JSEM Model	All Subsets Model (With Basement)	All Subsets Model (Without Basement)
<u>All Identified Variables (without higher degree and interaction effects)</u> No. of storey for podium (a), No. of storey for tower (b), No. of storey for basement (m), Square of no. of storey for podium (a^2), Square of no. of storey for tower (b^2), Average floor area for podium (f_p), Average floor area for tower (f_t), Average floor area for basement (f_b), Average storey height for podium (s_p), Average storey height for tower (s_t), Average storey height for basement (s_b), Average perimeter for tower and podium (p_{pt}), Average perimeter for basement (p_b), Roof area (r)	af_p, a^2f_p, bf_t, b^2f_t, abf_t, mf_b, $(as_p + bs_t)p_{pt}$, ms_bp_b, r (separating podium and tower)		
<u>Reduced Version of All Identified Variables (without higher degree and interaction effects)</u> No. of storey for superstructure (n), No. of storey for basement (m), Square of no. of storey for podium (n^2), Average floor area for superstructure (f_{pt}), Average floor area for basement (f_b), Average storey height for superstructure (s_{pt}), Average storey height for basement (s_b), Average perimeter for tower and podium (p_{pt}), Average perimeter for basement (p_b), Roof area (r)	nf_{pt}, n^2f_{pt}, mf_b, $ns_{pt}p_{pt}$, ms_bp_b, r (combining podium and tower)	n, m, n^2, f_{pt}, f_b, s_{pt}, s_b, p_{pt}, p_b, nf_{pt}, n^2f_{pt}, mf_b, ns_{pt}, ms_b, $n^2s_{p\,t}$, $ns_{pt}p_{pt}$, ms_bp_b, $n^2s_{pt}p_{pt}$, r	n, n^2, f_{pt}, s_{pt}, p_{pt}, nf_{pt}, n^2f_{pt}, ns_{pt}, n^2s_{pt}, $ns_{pt}p_{pt}$, $n^2s_{pt}p_{pt}$, r

Table 2: Included Candidates, Excluded Candidates and Selected Predictors for RJSEMs and RASEMs

RJSEM

	Office	Private Housing	Nursing Home	School
*afp / nfpt**	**o**	o	o	o
a2fp /	**o**	o	o	**o**
bft	**o**	**o**	NA	NA
b2ft	**o**	o	NA	NA
abft	o	o	NA	NA
mfb	o	o	o	NA
nsptppt	**o**	o	**o**	o
msbpb	x	x	o	NA
r	o	**o**	**o**	**o**

Legend:
o - Candidate x - Excluded Candidate
o - Selected Predictor NA - Not applicable

Remarks:
* - *afp* and *a2fp* for office and private housing, *nfpt* and *n2fpt* for nursing home and school

RASEM

	Office	Private Housing	Nursing Home	School
n	o	o	o	o
m	o	o	o	NA
n2	**o**	o	o	o
fpt	**o**	**o**	**o**	**o**
fb	o	**o**	o	NA
spt	o	**o**	o	o
sb	o	**o**	o	NA
ppt	**o**	o	o	o
pb	o	**o**	o	NA
nfpt	x	x	x	x
n2fpt	x	x	x	x
mfb	o	o	x	NA
nspt	**o**	o	o	**o**
msb	o	x	x	NA
n2spt	x	x	x	x
nsptppt	o	o	**o**	o
msbpb	x	x	x	NA
n2sptppt	x	x	x	x
r	x	x	x	x

Table 3: Summary of Means and Standard Deviations of Percentage Errors

	Office	Private Housing	Nursing Home	School
JSEM				
Mean % error (m)	-6.88%	-2.73%	2.09%	4.08%
SD of % error	21.43%	29.04%	20.03%	21.25%
MSQ of predicted price	1.19E+16	1.58E+16	2.32E+13	6.10E+12
p -value for *t* -test (H_0: m=0)	**0.04**	0.51	0.62	0.37
FAM				
Mean % error (m)	5.62%	1.31%	4.20%	3.35%
SD of % error	27.32%	23.53%	24.45%	21.45%
MSQ of predicted price	8.95E+15	1.25E+16	4.75E+13	6.70E+12
p -value for *t* -test (H_0: m=0)	0.19	0.69	0.42	0.46
CUBE				
Mean % error (m)	0.16%	1.47%	5.75%	3.56%
SD of % error	26.99%	19.59%	25.21%	24.56%
MSQ of predicted price	4.73E+15	1.05E+16	8.58E+13	8.88E+12
p -value for *t* -test (H_0: m=0)	0.97	0.60	0.29	0.49
RJSEM				
Mean % error (m)	3.06%	4.84%	3.21%	3.41%
SD of % error	25.38%	22.64%	21.45%	20.84%
MSQ of predicted price	6.26E+15	1.60E+16	2.73E+13	5.19E+12
p -value for *t* -test (H_0: m=0)	0.44	0.14	0.48	0.44
RASEM				
Mean % error (m)	2.96%	2.66%	3.09%	2.94%
SD of % error	22.15%	15.95%	21.36%	19.56%
MSQ of predicted price	4.99E+15	5.40E+15	2.85E+13	4.29E+12
p -value for *t* -test (H_0: m=0)	0.39	0.24	0.49	0.48

Remark:

Bold - *p* -value < 0.05, H_0 is rejected (i.e., Mean % error is significantly different from zero)

Table 4: Two-sample Mann-Whitney U-tests between Models for Offices and Private Housing

	Mann-Whitney *U*-test (at 99.17%* significance level)					
	Offices			Private Housing		
Pair	*Z*	*p*-value	H_0: No difference in absolute deviation (reject if p < 0.0083)	*Z*	*p*-value	H_0: No difference in absolute deviation (reject if p < 0.0083)
Common Comparisons for Both Groups						
JSEM and FAM	**-2.8896**	**0.0039**	**Reject H_0**	-1.8544	0.0637	Accept H_0
FAM and Cube	-1.3240	0.1855	Accept H_0	-1.6821	0.0926	Accept H_0
Cube and JSEM	-1.4493	0.1473	Accept H_0	**-3.0609**	**0.0022**	**Reject H_0**
Comparisons with RJSEM						
JSEM and RJSEM	-1.1988	0.2306	Accept H_0	-2.4818	0.0131	Accept H_0
FAM and RJSEM	-1.6103	0.1073	Accept H_0	-1.1651	0.2440	Accept H_0
Cube and RJSEM	-0.1252	0.9003	Accept H_0	-0.4481	0.6541	Accept H_0
Comparisons with RASEM						
JSEM and RASEM	-0.2952	0.7678	Accept H_0	**-4.3707**	**0.0000**	**Reject H_0**
FAM and RASEM	-2.2007	0.0278	Accept H_0	**-3.0126**	**0.0026**	**Reject H_0**
Cube and RASEM	-0.8946	0.3710	Accept H_0	-1.7441	0.0811	Accept H_0

Remark: * – 99.17% = (1 – 0.05/6) x 100%